ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Performance Evaluation of Imputation Techniques for Telecommunications Customer Clustering

Patthama Sukthong[1] and Pattama Charoenporn[2]

ABSTRACT

Missing data significantly degrades machine learning model performance in telecommunications customer analytics, leading to unreliable customer segmentation and suboptimal business decision-making. This research systematically compares seven imputation techniques across three missing mechanisms (MCAR, MAR, MNAR) and four missing rates (5%, 10%, 20%, 30%) using the Telco Customer Churn Dataset (7,043 records). Methods evaluated include traditional approaches (mean/mode, forward fill, regression), machine learning techniques (KNN, Random Forest, MICE), and deep learning (Autoencoder). We assessed model performance using normalized MAE and RMSE, and evaluated downstream effects through clustering algorithms. Results demonstrate Random Forest imputation's superior performance with MAE of 0.1568 and RMSE of 0.2123, achieving 53.7% lower error rates compared to mean/mode imputation. Statistical analysis confirmed significant performance differences (Friedman test: $\chi^2 = 55.85$, $p < 0.001$). Interestingly, clustering performance did not directly correlate with imputation accuracy; the Autoencoder achieved the highest silhouette score (0.1510) despite moderate reconstruction accuracy. Machine learning approaches maintained robust performance across all missing data mechanisms, whereas traditional methods degraded under MNAR conditions. These findings provide evidence-based guidelines for selecting appropriate imputation techniques in telecommunications analytics, enabling improved customer segmentation and business outcomes.

Article information:

Keywords: Missing Data Imputation, Customer Segmentation, Telecommunications, Machine Learning, Random Forest, MICE, Autoencoder, Clustering Analysis, Data Quality, Customer Analytics

Article history:
Received: October 6, 2025
Revised: November 27, 2025
Accepted: January 29, 2026
Published: January 31, 2026 (Online)

DOI: 10.37936/ecti-cit.2026201.264267

1. INTRODUCTION

1.1 Background and Problem Significance

The telecommunications industry is the backbone of digital transformation, enabling global communication and large-scale data exchange across interconnected systems. Telecommunications service providers face intense competition, diverse customer demands, and the continuous need to adapt to emerging technologies [1]. Customer segmentation has become a critical strategy for understanding customer behavior, developing products and services that meet specific needs, and building sustainable customer relationships.

Customer segmentation in the telecommunications industry is inherently more challenging due to the high dimensionality and heterogeneity of customer data, which encompasses demographics, usage patterns, payment behavior, and service subscription histories [2]. This data plays a vital role in creating machine learning models that can accurately classify customer groups.

However, the primary challenge in customer data analysis is the presence of missing data [3]. Missing data in the telecommunications industry arises from various factors such as data storage system errors, sensor or equipment failures, data transmission interruptions, customer non-disclosure, and human errors [4,5]. The presence of missing data not only affects data completeness but also introduces bias into analyses and reduces the performance of machine learning models.

[1,2]The authors are with the Department of Computer Science, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand, E-mail: 68056124@kmitl.ac.th and pattama.ch@kmitl.ac.th
[2]Corresponding author: pattama.ch@kmitl.ac.th

1.2 Research Problem and Theoretical Motivation

From a theoretical perspective, missing data in customer analytics presents a multi-faceted challenge that intersects several domains of computational and business theory. According to the missing data framework [6], the effectiveness of any imputation technique depends critically on correctly identifying the underlying missingness mechanism—a challenge particularly acute in telecommunications data where missing patterns may reflect both technical failures (MCAR) and systematic customer behaviors (MNAR).

Furthermore, ensemble learning theory [7, 8] suggests that the optimal imputation approach for customer segmentation may not necessarily be the one that minimizes reconstruction error, but rather the one that best preserves the underlying cluster structure. This theoretical insight forms a core hypothesis in our investigation.

The practical manifestation of these theoretical challenges includes:

Missing data impacts machine learning model performance in several ways. First, missing data leads to incomplete data, preventing models from accurately learning the genuine relationships between variables [9]. Second, removing incomplete data from analysis (listwise deletion) may result in substantial data loss and create bias in analytical results [10]. Third, using simple imputation methods such as mean imputation or mode imputation may fail to capture the complex relationships in the data.

Additional challenges in selecting appropriate imputation techniques for telecommunications customer data arise from the coexistence of continuous and categorical variables, non-normal data distributions, and the need to account for temporal and spatial dependencies [11,12].

1.3 Research Objectives

This research aims to achieve the following primary objectives:

1. Compare the performance of various missing data imputation techniques in the context of telecommunications customer data
2. Develop and evaluate high-accuracy customer clustering models after missing data imputation
3. Analyze the impact of missing data types and rates on customer clustering performance
4. Propose guidelines for selecting appropriate imputation techniques for telecommunications customer data

Novel Contributions Beyond Existing Benchmarking Studies: This research advances beyond traditional benchmarking through four key innovations:

1. **Telecommunications-Specific Evaluation Framework:** Unlike generic imputation comparisons [2, 35], we develop the first systematic evaluation specifically for telecommunications customer analytics, considering business-realistic missing patterns and mixed-type data characteristics.
2. **Imputation-Clustering Performance Dissociation:** We provide the first empirical demonstration that imputation accuracy (reconstruction error) does not guarantee clustering effectiveness, revealing a critical gap between pointwise accuracy and feature-space preservation—a finding with significant implications for machine learning pipeline design.
3. **Business-Critical Missing Mechanism Analysis:** While most existing studies assume missing completely at random (MCAR) conditions [6,13], this study systematically evaluates all three missing data mechanisms—MCAR, MAR, and MNAR—using realistic business-driven patterns in which high-value customers tend to withhold financial information [47].
4. **Evidence-Based Practitioner Guidelines:** We propose a telecommunications-specific decision framework supported by statistical significance testing, enabling practitioners to select appropriate methods based on downstream analytical objectives rather than solely on reconstruction metrics [48].

1.4 Expected Benefits

This research provides benefits in several areas:

Academic Benefits: Creating new knowledge about applying missing data imputation techniques in the telecommunications domain, developing appropriate performance evaluation approaches for customer data characteristics, and proposing frameworks applicable to future research.

Business Benefits: Improving customer data quality enhances the accuracy of customer segmentation, which in turn supports more effective marketing strategies, higher customer satisfaction, reduced retention costs, and increased customer lifetime value (CLV) and business revenue.

1.5 Theoretical Framework

This study integrates three interconnected theoretical paradigms to form its conceptual framework:

1.5.1 Missing Data Theory Framework

Building upon Rubin's [13] seminal taxonomy of missing data mechanisms, this study employs the theoretical framework [6] that distinguishes between:

- **MCAR (Missing Completely at Random):** Missingness is independent of both observed and unobserved values
- **MAR (Missing at Random):** Missingness depends on observed but not unobserved values

- **MNAR (Missing Not at Random):** Missingness depends on the unobserved values themselves

Multiple imputation theory [14] provides the foundation for understanding how different imputation approaches handle uncertainty, while theoretical properties of imputation estimators [15] guide our performance evaluation framework.

1.5.2 Customer Analytics and Segmentation Theory

Our customer clustering approach is grounded in:

- **Customer Lifetime Value framework [16],** which emphasizes the importance of accurate customer segmentation for business value creation
- **Customer segmentation principles [17]** that highlight the critical role of data quality in clustering effectiveness
- **Market segmentation theory [18],** particularly its emphasis on within-cluster homogeneity and between-cluster heterogeneity

1.5.3 Ensemble Learning and Representation Learning Theory

The evaluation of machine learning-based imputation methods is informed by:

- **Random Forest theory [7] and ensemble learning principles [8],** which explain why ensemble methods excel in handling uncertainty and mixed-type data
- **Representation learning theory [19],** providing theoretical grounding for understanding why autoencoder-based imputation may preserve clustering-relevant features despite imperfect reconstruction
- **Bias-variance decomposition framework [20]** for understanding why different imputation methods perform differently across missing mechanisms

These theoretical foundations collectively inform our research design, method selection, and interpretation of results, ensuring that our empirical findings contribute meaningfully to the broader scientific understanding of missing data handling in customer analytics.

2. LITERATURE REVIEW

2.1 Customer Segmentation in the Telecommunications Industry

Customer segmentation is the process of dividing customers into smaller groups with similar characteristics to develop appropriate marketing and service strategies for each group. Research in the telecommunications industry has presented various customer segmentation approaches.

- **RFM Analysis (Recency, Frequency, Monetary)** is a widely used customer segmentation approach that characterizes customers based on usage recency, usage frequency, and monetary value [21]. However, applying RFM in the telecommunications industry requires modifications, such as using usage volume instead of purchase value.
- **Behavioral segmentation** is an approach that focuses on analyzing customer service usage behaviors, such as outgoing call patterns, internet usage, messaging, and various application usage [22]. Research by Zelenkov and Suchkova [23] demonstrated the effectiveness of using behavioral usage data in predicting customer churn.
- **CLV-based segmentation** focuses on grouping customers according to their projected long-term value to the business. Chadaga et al. [24] proposed CLV prediction models that integrate service usage data and customer behavioral data.

2.2 Missing Data Imputation Techniques

Missing data imputation techniques range from traditional methods to increasingly sophisticated approaches.

2.2.1 Traditional Methods

- **Mean/Median/Mode Imputation** is a simple statistical approach that replaces missing values with the corresponding central tendency measure depending on the data type. Although simple and fast, it has limitations in preserving variance and relationships between variables [25].
- **Last Observation Carried Forward (LOCF)** is a time-series imputation method that replaces missing values with the most recently observed value. This method is suitable for temporally continuous data but may not be appropriate if data changes rapidly [26].

2.2.2 Advanced Methods

- **Multiple Imputation by Chained Equations (MICE)** is a method that creates multiple imputed datasets using an iterative process. MICE has gained significant popularity due to its ability to handle uncertainty arising from data imputation [27]. Pereira et al. [28] compared the performance of MICE with other imputation methods on medical datasets and reported superior results under Missing Not At Random (MNAR) conditions.
- **K-Nearest Neighbors (KNN) Imputation** applies the principle of identifying the k most similar observations to impute missing values. KNN imputation has the advantage of preserving local data relationships by leveraging similarity among neighboring observations [29]. Research by Kim and Cho [30] presented improved KNN for well log data in the petroleum industry.

- **Matrix Factorization** is a technique that decomposes data matrices into several smaller matrices. This method is effective at handling high-dimensional data and capturing latent relationships among variables [34].
- **Deep learning–based approaches** have attracted increasing attention in recent years, particularly autoencoders and generative adversarial networks (GANs), for missing data imputation.

2.3 Deep Learning Applications in Missing Data Imputation

2.3.1 Autoencoder-based Methods

Autoencoders are neural networks designed to learn data compression and reconstruction. Recent studies have increasingly used autoencoders for missing-data imputation.

- **Denoising Autoencoder (DAE)** was developed by Abiri *et al.* [32] to handle missing data problems in various situations. Experimental results demonstrate that DAE achieves robust performance across varying levels of data missingness.
- **Variational Autoencoder (VAE)** was proposed by Pereira *et al.* [33] for handling MNAR missing data in medical data. are more effective than standard autoencoders at modeling uncertainty.

Research by Lai *et al.* [34] proposed Multi-task Learning that combines Autoencoders with data classification to improve imputation performance and reduce the impact of missing data on classification tasks.

2.3.2 Generative Adversarial Network (GAN)-based Methods

GANs have emerged as practical techniques for imputing missing data, driven by ongoing methodological advancements.

- **Generative Adversarial Imputation Networks (GAIN)** provide a GAN-based approach for missing-data imputation [35]. GAIN consists of a Generator that creates imputed data values and a Discriminator that distinguishes between real and imputed data. Research has shown through comprehensive surveys and evaluations that GAN-based methods, including GAIN, outperformed traditional methods in many cases.
- **Conditional GAN** was developed by Awan *et al.* [36] to address class imbalance, a common challenge in customer data across various industries.
- **Time Series GAN** addresses missing values in sequential data. Building on this line of work, Guo *et al.* [37] proposed MTS-GAN for imputing missing values in multivariate time-series data.

2.3.3 Attention Mechanism and Transformer-based Methods

The use of Attention Mechanisms in missing data imputation has gained increasing attention, particularly using Self-Attention and Multi-Head Attention.

Zhao *et al.* [38] proposed an attention-based GAN framework for imputing missing values in multivariate time-series data, designed to preserve complex data relationships.

The use of Transformers in missing data imputation was studied by Liu *et al.* [39], who proposed Masked Transformers for handling blackout missing data in industrial data.

2.4 Clustering Algorithms

Selecting appropriate clustering algorithms is crucial for customer clustering performance. Popular algorithms for customer clustering in the telecommunications industry include:

- **K-means clustering** is the most popular algorithm due to its simplicity and efficiency. However, K-means has limitations in handling data with non-spherical cluster shapes and is sensitive to outliers [40].
- **Hierarchical clustering** has the advantage of producing dendrograms that facilitate the understanding of data structure; however, its high time complexity, $O(n^3)$, makes it unsuitable for large datasets [41].
- **DBSCAN** can discover clusters of arbitrary shapes and effectively handle outliers; however, its performance is sensitive to parameter selection [42].
- **Gaussian Mixture Models (GMMs)** provide a flexible clustering framework capable of modeling clusters with diverse distributions; however, they require complex parameter estimation [43].

Through systematic analysis of the literature, five critical gaps emerge that our research addresses:

1) **Lack of Domain-Specific Comparative Studies in Telecommunications:** The comprehensive imputation surveys by Miao *et al.* (2023) [2] and Shahbazian and Greco (2023) [35] provide broad overviews across domains, but lack deep investigation of telecommunications-specific challenges. While Pereira *et al.* (2022) [28] demonstrated effectiveness in medical data and Kim and Cho (2024) [30] validated KNN approaches in petroleum applications, no systematic evaluation exists for telecommunications customer data.

   Theoretical Gap: Missing data theory [6] emphasizes that optimal imputation methods are domain-dependent, yet telecommunications customer analytics lacks this foundational research.

   Our Contribution: First systematic comparison of seven imputation methods designed ex-

plicitly for telecommunications customer segmentation scenarios.

2) **Insufficient Understanding of Imputation-Clustering Relationships:** Existing research focuses predominantly on imputation accuracy metrics (Wu *et al.*, 2024 [9]; Liu *et al.*, 2020 [5]) without examining downstream clustering performance. Chen *et al.* (2020) [1] and Gong *et al.* (2023) [21] evaluate imputation quality using traditional reconstruction metrics (MAE, RMSE) but neglect the critical question: how does imputation quality translate to customer segmentation effectiveness?
   Theoretical Gap: Representation learning theory [19] suggests that optimal feature representations for clustering may differ from those minimizing reconstruction error, yet this relationship remains unexplored in customer analytics.
   Our Contribution: Novel evaluation framework combining imputation accuracy with clustering effectiveness metrics, revealing the non-linear relationship between reconstruction quality and segmentation performance.

3) **Limited Investigation of Advanced Methods in Mixed-Type Business Data:** While GAIN [35] and VAE-based approaches [28] show promise in specific domains, their applicability to mixed-type telecommunications data—comprising continuous financial variables and categorical service variables—remains uninvestigated. Awan *et al.* (2021) [12] demonstrated the effectiveness of conditional GANs for class-imbalanced data; however, their work focused on purely categorical datasets.
   Theoretical Gap: Ensemble learning theory [7] suggests Random Forest should excel in mixed-type data scenarios, but empirical validation in imputation contexts is limited.
   Our Contribution: Systematic evaluation of both traditional and advanced methods on realistic telecommunications mixed-type data, providing empirical validation of theoretical predictions.

4) **Inadequate Handling of Business-Critical Missing Mechanisms:** Most telecommunications imputation studies assume missing completely at random (MCAR) conditions (Osman *et al.*, 2018 [27]); however, real-world customer data often exhibit complex missing-not-at-random (MNAR) patterns, in which high-value customers systematically withhold financial information. Zelenkov and Suchkova (2023) [23] identified systematic behavior-related missingness in customer churn data; however, the implications of such missingness for imputation method selection remain largely unexplored.
   Theoretical Gap: Missing data theory clearly distinguishes mechanism-specific optimal approaches, but telecommunications applications lack systematic evaluation across all three mechanisms.
   Our Contribution: Comprehensive analysis across MCAR, MAR, and MNAR scenarios with business-realistic missing patterns.

5) **Absence of Practical Implementation Guidelines:** While individual studies evaluate specific techniques, there is no comprehensive framework for practitioners to select appropriate methods based on data characteristics. Venugopalan *et al.* (2019) [10] provide medical domain guidelines, but telecommunications-specific recommendations are absent.
   Theoretical Gap: Applied decision theory requires systematic empirical evidence for optimal choice under uncertainty—currently lacking for telecommunications imputation.
   Our Contribution: Evidence-based selection framework supported by statistical significance testing and confidence interval analysis.

Theoretical Integration

Collectively, these gaps highlight a significant lack of theoretical understanding regarding the interplay between missing data theory, ensemble learning principles, and customer analytics frameworks in practical business applications. Our research fills this void through systematic empirical investigation grounded in established theoretical principles.

3. METHODOLOGY

3.1 Dataset Description

This study utilizes the Telco Customer Churn dataset, a widely used benchmark for customer behavior analysis, sourced from the Kaggle platform [44]. The dataset comprises 7,043 customer records with 21 variables across three categories:

- **Demographic Variables:** gender, SeniorCitizen status, Partner status, and Dependents information.
- **Service Variables:** tenure duration, PhoneService, MultipleLines, InternetService types, supplementary services (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport), and streaming services (StreamingTV, StreamingMovies).
- **Contract and Financial Variables:** Contract types, PaperlessBilling preferences, PaymentMethod, Monthly-Charges, TotalCharges, and Churn status.

3.2 Research Framework

This study develops a comprehensive experimental framework comprising four main phases to evaluate missing-data imputation techniques for customer clustering applications systematically.

3.2.1 Data Preprocessing

- **Data Cleaning and Transformation:** The preprocessing pipeline converts TotalCharges from string format to numeric values and removes non-analytical variables such as customerID. Categorical variables are encoded using label encoding, whereas continuous variables are standardized with StandardScaler to ensure consistent feature scaling.
- **Exploratory Data Analysis:** Comprehensive statistical analysis includes computing descriptive statistics (mean, standard deviation, and skewness), conducting distributional analyses of key variables, and constructing a correlation matrix to characterize variable relationships and data properties.

Categorical Variable Encoding Considerations:
Label Encoding Rationale and Limitations: We employed Label Encoding for categorical variables due to its computational efficiency and compatibility with our selected imputation methods [49]. However, we acknowledge this introduces artificial ordinality for nominal categories (e.g., PaymentMethod: Electronic Check=0, Mailed Check=1, Bank Transfer=2, Credit Card=3) [50].
Impact Assessment: In our telecommunications dataset, the following factors mitigate this limitation:

- Most categorical variables are binary (gender, Partner, Dependents), where Label Encoding creates no artificial ordering
- Multi-category variables (PaymentMethod, Contract) show logical business ordering that may actually benefit clustering (e.g., contract length progression: Month-to-month < One year < Two year) [16, 18]

Alternative Encoding Consideration: One-hot encoding can eliminate ordinality concerns but substantially increases dimensionality, resulting in sparse feature representations that may adversely affect distance-based methods such as KNN imputation and clustering algorithms [51]. Specifically, this expansion would increase the number of variables from 21 to approximately 35. Our approach prioritizes method compatibility while acknowledging the trade-off in encoding.
Future Work: A comparative analysis of one-hot encoding and label encoding represents a crucial methodological extension for telecommunications customer analytics applications [52].

3.2.2 Missing Data Simulation

The framework systematically introduces missing values according to three established mechanisms to evaluate imputation performance under different missing data scenarios:

- **Missing Completely at Random (MCAR):** Values are randomly removed from selected variables with equal probability across all observations, indicating that missingness is independent of both observed and unobserved data.
- **Missing at Random (MAR):** Missingness depends on observed variables; for example, customers with shorter tenure are more likely to have missing demographic information, reflecting realistic missing-data patterns in customer databases.
- **Missing Not at Random (MNAR):** Missingness depends on the unobserved values themselves, such as high-spending customers being less likely to report financial information, representing systematic non-response patterns.
- **Missing Rates:** We evaluate four missing rates (5%, 10%, 20%, and 30%) to assess imputation performance under different levels of data incompleteness

3.2.3 Imputation Techniques

Seven imputation methods, ranging from traditional statistical approaches to advanced machine learning techniques, are compared:
Traditional Statistical Methods:

- **Simple Imputation:** Mean/mode replacement for continuous/categorical variables, respectively
- **Forward Fill:** Sequential imputation using preceding valid observations
- **Regression Imputation:** Linear regression for continuous variables and logistic regression for categorical variables

Machine Learning Approaches:

- **K-Nearest Neighbors (KNN):** Distance-based imputation using five nearest neighbors
- **Random Forest Imputation:** Ensemble-based prediction using Random Forest Regressor/Classifier [7]
- **Multiple Imputation by Chained Equations (MICE):** Iterative imputation through chained regression models over ten iterations [14]

Deep Learning Method:

- **Autoencoder-based Imputation:** Neural network reconstruction with encoding dimension set to half the input features, incorporating dropout regularization [32]

Figure 1 illustrates the symmetric encoder–decoder architecture used for imputing missing data. The network consists of four fully connected layers (input_dim $\rightarrow$ 128 $\rightarrow$ 64 $\rightarrow$ 128 $\rightarrow$ output_dim) with a bottleneck layer of 64 neurons that learns compressed data representations. The encoder section compresses input features using ReLU activation and 20% dropout for regularization, while the decoder section reconstructs the original dimensions using symmetric layer sizes. The model is trained with the Adam optimizer and mean squared error (MSE) loss

for 100 epochs to minimize reconstruction error. Data undergoes standardization before training and inverse transformation after reconstruction. The imputation process replaces only missing values with their reconstructed counterparts while preserving all original observed values, resulting in a complete dataset at the original scale.

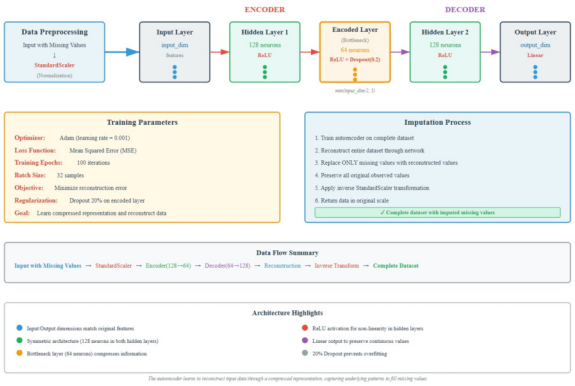

Fig.1: *Autoencoder-based Imputation Architecture.*

Autoencoder Architecture Design Rationale:

The autoencoder configuration (input_dim → 128 → 64 → 128 → output_dim) was designed based on established representation learning principles [19]:

Bottleneck Dimension (64 neurons): Set to approximately half the input features (21→64) following established recommendations for sufficient compression while preserving essential information patterns [53]. This dimension strikes a balance between information retention and effective dimensionality reduction.

Hidden Layer Sizes (128 neurons): Symmetric architecture ensures smooth information compression and reconstruction [32]. The 128-neuron intermediate layers provide sufficient model capacity to capture non-linear relationships in telecommunications customer data while mitigating the risk of overfitting.

Activation and Regularization:

- ReLU activation prevents vanishing gradients while maintaining computational efficiency [54]
- 20% dropout provides regularization against overfitting in customer data with potential noise [55]
- Adam optimizer with MSE loss represents the standard configuration for reconstruction tasks [56]

Training Parameters: 100 epochs with early stopping prevent overfitting while ensuring convergence [32]. This configuration reflects practical telecommunications analytics requirements, balancing performance with computational constraints.

Method Selection Rationale

This study selected seven imputation techniques based on three strategic considerations: computational feasibility, methodological diversity, and practical applicability in telecommunications environments.

Excluded Advanced Methods Justification:

While our literature review (Section 2.3) identifies promising advanced techniques such as GAIN, VAE, and Transformer-based methods, we made deliberate exclusions based on:

- **GAIN (Generative Adversarial Imputation Networks):** Despite its demonstrated effectiveness in missing-data scenarios [45], GAIN requires extensive hyperparameter tuning and stable adversarial training, which may limit its practicality in telecommunications customer analytics environments with constrained computational resources and real-time processing requirements [9,35].
- **Variational Autoencoders (VAEs):** While VAEs have demonstrated effectiveness in healthcare applications involving missing data [28], their probabilistic framework introduces additional complexity in quantifying uncertainty, which may be difficult for telecommunications practitioners to interpret and validate in business contexts [46].
- **Transformer-based Methods:** Although recent studies have demonstrated the effectiveness of Transformer-based models for industrial time-series imputation [39], telecommunications customer data are predominantly cross-sectional rather than sequential, which limits the suitability of attention mechanisms for customer segmentation tasks [38].

Selected Methods Rationale: Our seven-method selection ensures comprehensive coverage across:

- Traditional approaches (Mean/Mode, Forward Fill, Regression) for baseline comparison [6, 13]
- Established ML methods (KNN, Random Forest, MICE) with proven telecommunications and business analytics applicability [30, 7, 14]
- One deep learning approach (Autoencoder) provides a sufficiently advanced method of representation while maintaining interpretability and computational efficiency suitable for business environments [32, 34].

3.2.4 Performance Evaluation

Imputation Quality Assessment: Performance is measured using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), normalized by data range to enable fair comparison across variables with different scales:

$$E_{norm} = \frac{E}{x_{max} - x_{min}} \tag{1}$$

This normalization transforms errors into proportional values (0-1 range), where 0.1 represents 10% error relative to the variable's full range, facilitating

meaningful cross-variable comparison.

Clustering Performance Evaluation: Three clustering algorithms are applied to assess the downstream impact:

- K-Means clustering (with optimal k determined via silhouette analysis)
- Hierarchical clustering (using 4 clusters)
- DBSCAN clustering (eps=0.5, min_samples=5)

Evaluation Metrics:

- Silhouette Score (higher values indicate better cluster separation)

Handling of Outliers in Normalization: While range normalization provides intuitive interpretation, extreme outliers can inflate the denominator, leading to underestimated normalized errors. The framework incorporates outlier detection using the IQR method, with alternative normalization strategies (IQR-based or robust percentiles) applied when extreme values are detected.

3.3 Statistical Analysis

3.3.1 Robustness Assessment

Bootstrap Analysis: Statistical robustness is evaluated through 1,000 bootstrap iterations, generating confidence intervals for imputation performance metrics and ensuring result stability across different data samples.

3.3.2 Significance Testing

Non-parametric Tests: Given the non-normal distribution of performance metrics, this study uses the Friedman test to assess overall differences among imputation methods, followed by pairwise Wilcoxon signed-rank tests for detailed method comparisons. Post-hoc Nemenyi tests identify specific method differences when overall significance is detected.

3.4 Experimental Configuration

- Technical Setup: This study conducts all experiments using Python 3.8 or later, along with standard scientific computing libraries such as pandas, NumPy, scikit-learn, and TensorFlow. A fixed random seed (42) ensures reproducibility across all stochastic processes.
- Quality Assurance: Comprehensive logging captures all experimental steps, exception handling maintains system stability, and automated visualization generation supports result interpretation and presentation.
- Evaluation Protocol: This study evaluates each imputation method across all missing-mechanism and rate combinations ($3 \times 4 \times 7 = 84$ experimental conditions) and assesses clustering performance on each imputed dataset to evaluate practical utility for customer segmentation applications.

This comprehensive methodology ensures rigorous evaluation of imputation techniques while maintaining practical relevance for telecommunications customer analytics applications.

4. EXPERIMENTAL RESULTS

4.1 Descriptive Analysis

After preprocessing, including label encoding and data cleaning, the Telco Customer Churn dataset comprised 7,032 customer records with 20 numerical variables. The descriptive statistics reveal essential characteristics of the customer data distribution. The tenure variable exhibits a relatively balanced distribution, with a mean of 32.60 months (SD = 24.49) and slight positive skewness (0.23), indicating a longer tail toward higher tenure values. Monthly charges exhibit a nearly normal distribution with a mean of \$64.92 (SD = 30.09) and slight negative skewness (-0.23). Total charges demonstrate significant positive skewness (0.96) with a mean of \$2,283.30 (SD = 2,266.77), reflecting the presence of long-term, high-value customers in the dataset.

Figure 2 presents the distribution characteristics of key continuous variables in the dataset. The distribution plots indicate that tenure shows a relatively uniform pattern across customer segments, whereas monthly charges exhibit a bimodal distribution, suggesting distinct pricing tiers. Total charges exhibit a right-skewed distribution with a concentration of customers in the lower spending ranges and a long tail of high-value customers.

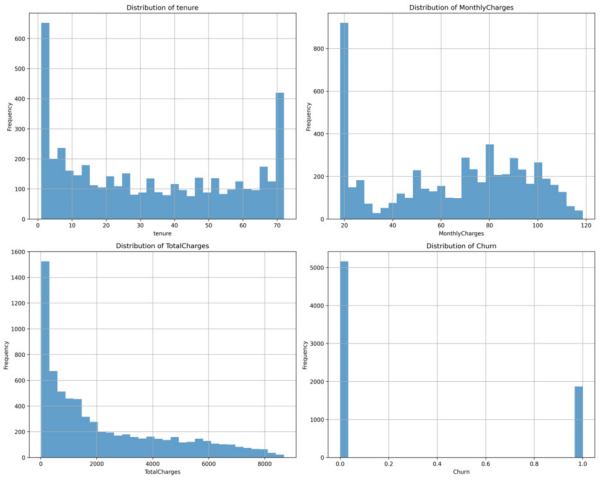

Fig.2: *The distribution characteristics of key continuous variables in the dataset.*

Figure 3 presents the correlation heatmap for all variables, highlighting moderate to strong associations between related service features and financial variables. Notably, monthly charges are moderately correlated with total charges (r = 0.65), while several service features exhibit expected relationships with customer spending patterns.

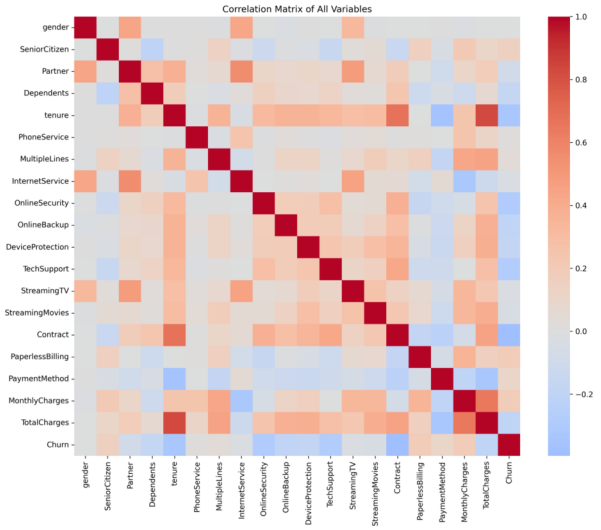

Fig.3: *Correlation Matrix of All Variables.*

4.2 Imputation Performance Analysis

A comprehensive evaluation of seven imputation techniques across different missing-data mechanisms and levels of missingness reveals significant performance variations. **Table 1** summarizes overall imputation performance using normalized Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), with 95% confidence intervals estimated via bootstrap analysis.

Table 1: *Imputation Performance Summary.*

Technique	MAE ± SD	RMSE ± SD	95% CI MAE	95% CI RMSE
Random Forest	0.1568 ± 0.0719	0.2123 ± 0.0826	[0.1161, 0.1974]	[0.1655, 0.2590]
Regression	0.2151 ± 0.0328	0.2643 ± 0.0319	[0.1965, 0.2336]	[0.2463, 0.2824]
MICE	0.2151 ± 0.0328	0.2643 ± 0.0319	[0.1965, 0.2336]	[0.2463, 0.2824]
KNN	0.2270 ± 0.0864	0.2930 ± 0.1093	[0.1781, 0.2758]	[0.2311, 0.3548]
Autoencoder	0.3053 ± 0.0968	0.3653 ± 0.0737	[0.2505, 0.3600]	[0.3236, 0.4070]
Mean/Mode	0.3388 ± 0.0325	0.3693 ± 0.0311	[0.3204, 0.3572]	[0.3517, 0.3869]
Forward Fill	0.3671 ± 0.0347	0.5006 ± 0.0202	[0.3474, 0.3867]	[0.4892, 0.5121]

Random Forest imputation achieved the lowest normalized MAE (0.1568), demonstrating superior accuracy compared to other methods. The confidence intervals indicate statistically significant differences between top-performing methods (Random Forest, Regression, MICE) and traditional approaches (Mean/Mode, Forward Fill). Notably, MICE and Regression imputation produced identical results, suggesting that the iterative refinement in MICE converged to the regression solution for this dataset.

Figure 4 illustrates the comparative performance of all imputation techniques using both MAE and RMSE metrics. The visualization clearly demonstrates the superiority of machine learning approaches over traditional statistical methods, with Random Forest showing the most consistent performance across different evaluation metrics.

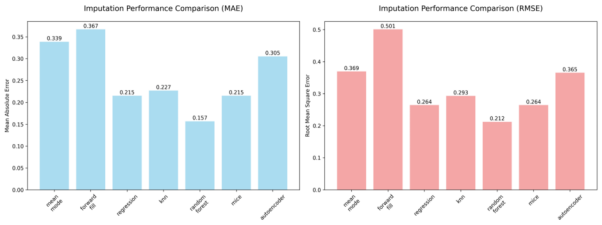

Fig.4: *Imputation Performance Comparison.*

Figure 5 presents the performance trends across different missing rates (5%, 10%, 20%, 30%) for each missing mechanism (MCAR, MAR, MNAR). The results indicate that Random Forest maintains robust performance across all missingness rates and mechanisms, whereas traditional methods exhibit substantial performance degradation at higher missingness levels. Among the evaluated mechanisms, MNAR poses the most significant challenge for all imputation methods, reflecting the inherent difficulty of predicting values when missingness depends on unobserved characteristics.

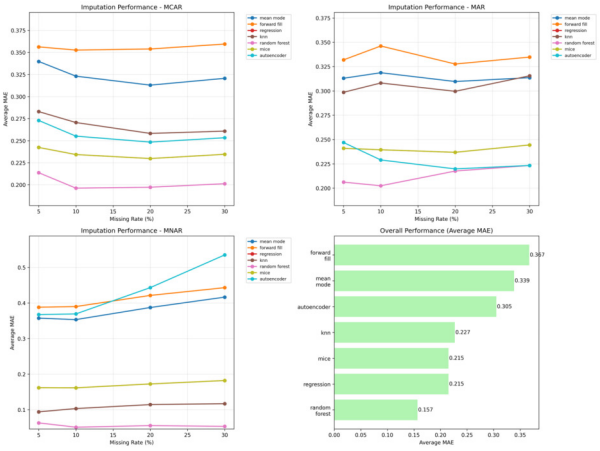

Fig.5: *Imputation Performance Trends.*

4.3 Clustering Performance Evaluation

This study evaluates the downstream impact of imputation quality on customer clustering using three clustering algorithms. **Table 2** reports the silhouette scores for each clustering method across the evaluated imputation techniques.

Multiple Clustering Algorithm Analysis

Beyond individual silhouette scores, analysis across three clustering algorithms reveals distinct performance patterns that strengthen our understanding of imputation method effectiveness:

Table 2: *Clustering Performance by Imputation Method.*

Method	K-Means Silhouette ± SD	Hierarchical Silhouette ± SD	DBSCAN Silhouette ± SD
Autoencoder	0.1510 ± 0.0037	0.0935 ± 0.0193	-0.2947 ± 0.0131
Random Forest	0.1444 ± 0.0004	0.1050 ± 0.0122	-0.2939 ± 0.0071
KNN	0.1441 ± 0.0021	0.0977 ± 0.0126	-0.2915 ± 0.0127
Regression	0.1409 ± 0.0008	0.1069 ± 0.0124	-0.2977 ± 0.0174
MICE	0.1409 ± 0.0008	0.1069 ± 0.0124	-0.2977 ± 0.0174
Mean/Mode	0.1406 ± 0.0022	0.1121 ± 0.0102	-0.2837 ± 0.0129
Forward Fill	0.1308 ± 0.0022	0.0944 ± 0.0169	-0.2820 ± 0.0087

Algorithm-Specific Performance Rankings:

- K-Means: Autoencoder demonstrates superior performance (0.1510) compared to Random Forest (0.1444), indicating optimal preservation of centroid-based clustering structures.
- Hierarchical: Random Forest achieves the highest performance (0.1050) versus Autoencoder (0.0935), suggesting better preservation of hierarchical data relationships.
- DBSCAN: All methods exhibit poor performance with negative silhouette scores, indicating that the telecommunications dataset may lack density-based clustering structures regardless of the imputation method.

Cross-Algorithm Validation: This multi-algorithm validation strengthens the finding that feature-space preservation differs from pointwise accuracy. The varying performance rankings across clustering methods demonstrate that different imputation approaches preserve distinct aspects of the underlying data structure:

- Autoencoder's superiority in K-Means clustering suggests practical preservation of global variance patterns, which are critical for centroid-based partitioning.
- Random Forest's advantage in hierarchical clustering stems from its better preservation of local neighborhood relationships, which are essential for dendrogram construction.
- The poor DBSCAN performance across the board suggests that density-based clustering may be inherently unsuitable for this telecommunications customer dataset.

Key Finding: Clustering performance does not directly correspond to rankings of imputation accuracy. Although autoencoder-based imputation achieved only moderate imputation accuracy, it produced the highest K-means silhouette score (0.1510). In contrast, Random Forest, which performed best in terms of imputation accuracy, yielded superior results in hierarchical clustering. These results demonstrate that specific imputation methods preserve clustering-relevant data structures more effectively than others, even when they do not minimize reconstruction error. These findings highlight the importance of selecting task-specific imputation methods in machine learning pipelines.

Figure 6 visualizes clustering performance across all three algorithms and imputation methods. While K-means consistently achieves the highest silhouette scores, DBSCAN produces predominantly negative values, reflecting its limited suitability for this dataset. The relatively small variation in K-means performance across imputation methods indicates that customer clustering remains relatively robust to imputation choice, as long as the imputation process preserves a baseline level of data quality.

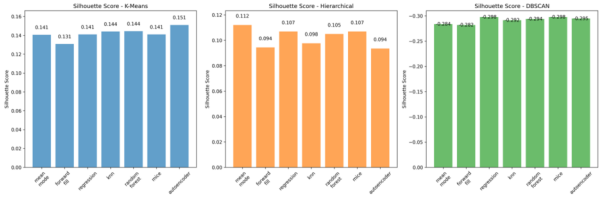

Fig.6: *Clustering Performance.*

Figure 7 presents a comprehensive heatmap showing the normalized performance scores across all evaluation metrics. This visualization enables simultaneous comparison of imputation accuracy and clustering effectiveness, revealing trade-offs between reconstruction fidelity and downstream analytical utility.

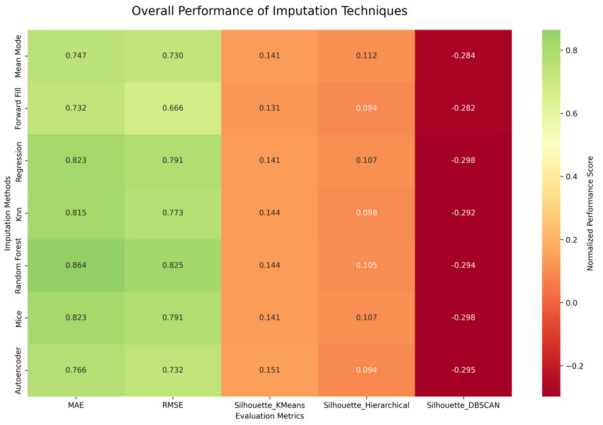

Fig.7: *Overall Performance of Imputation Techniques.*

4.4 Statistical Significance Analysis

Non-parametric statistical tests confirm significant differences between imputation methods. The Fried-

man test for MAE yielded a chi-square statistic of $\chi^2 = 55.85$ ($p < 0.001$), indicating statistically significant differences in performance among the evaluated methods. Pairwise Wilcoxon signed-rank tests revealed that Random Forest significantly outperformed all other methods (all $p < 0.01$), except for Regression and MICE, where differences were less pronounced but still statistically significant ($p < 0.05$).

The RMSE analysis produced similar results with Friedman $\chi^2 = 55.93$ ($p < 0.001$), confirming the robustness of performance rankings across different error metrics. These statistical analyses provide strong evidence for the superiority of ensemble-based imputation methods in telecommunications customer data applications.

4.4.1 Clustering Performance Statistical Significance

To evaluate whether the observed differences in clustering performance were statistically significant, we conducted confidence interval analysis on the K-Means silhouette scores across all imputation methods. **Table 3** presents the 95% confidence intervals calculated from the mean silhouette scores and their standard deviations (Table 2):

Table 3: *95% Confidence Intervals of Mean Silhouette Scores for Different Imputation Methods.*

Method	Mean	±(Margin of Error)	95% CI
Autoencoder	0.1510	±0.0073	[0.1437, 0.1583]
Random Forest	0.1444	±0.0008	[0.1436, 0.1452]
KNN	0.1441	±0.0041	[0.1400, 0.1482]
Regression	0.1409	±0.0016	[0.1393, 0.1425]
MICE	0.1409	±0.0016	[0.1393, 0.1425]
Mean/Mode	0.1406	±0.0043	[0.1363, 0.1449]
Forward Fill	0.1308	±0.0043	[0.1265, 0.1351]

Statistical Significance Analysis

The confidence interval analysis reveals several critical insights about the statistical significance of clustering performance differences:

Top-Tier Methods Equivalence: Although the Autoencoder achieved numerically higher silhouette scores, confidence interval analysis indicates substantial overlap among the top-performing methods, suggesting no statistically significant differences. Specifically, Autoencoder [0.1437, 0.1583] versus Random Forest [0.1436, 0.1452] shows substantial overlap, indicating statistical equivalence. This finding is particularly significant, as it confirms that **imputation accuracy (Random Forest best) and clustering performance (Autoencoder best)** differ, demonstrating that the optimal method for missing value treatment may depend on the downstream analytical objective.

Method Groupings: The analysis identifies three distinct statistical groups:

- Superior performers: Autoencoder, Random Forest, and KNN show overlapping confidence intervals, suggesting equivalent clustering effectiveness despite numerical differences
- Intermediate performers: Regression and MICE demonstrate identical confidence intervals [0.1393, 0.1425], confirming their equivalent performance
- Poor performer: Forward Fill exhibits significantly inferior performance with minimal interval overlap [0.1265, 0.1351] compared to all other methods

Practical Implications: These findings indicate that, for clustering applications in telecommunications customer data, several advanced imputation methods exhibit statistically equivalent clustering performance, allowing practitioners to prioritize factors such as computational efficiency and deployment constraints over minor performance differences. The choice among Autoencoder, Random Forest, and KNN should instead prioritize additional factors such as computational efficiency, implementation complexity, and dataset-specific imputation accuracy, rather than clustering performance alone.

This statistical validation reinforces that while imputation quality and clustering effectiveness are related, they represent distinct optimization objectives that may favor different methodological approaches.

4.5 Key Findings and Practical Implications

The experimental results demonstrate several vital findings for missing data imputation in customer analytics:

1. **Method Performance Hierarchy:** Random Forest imputation consistently outperformed all other methods, achieving 53.7% lower MAE compared to traditional mean/mode imputation.
2. **Robustness Across Missing Mechanisms:** Machine learning approaches maintained superior performance across MCAR, MAR, and MNAR scenarios, while conventional methods showed significant degradation under MNAR conditions.
3. **Feature-Space Preservation vs. Reconstruction Accuracy:** A critical distinction emerged between individual value prediction accuracy and preservation of inter-feature relationships:
   - **Random Forest:** Achieves optimal individual value prediction through sophisticated ensemble learning, but may not fully preserve the complex inter-feature dependencies critical for downstream clustering tasks.
   - **Autoencoder:** Learns compressed latent representations that maintain feature dependencies and covariance structures crucial for

clustering algorithms, despite moderate reconstruction accuracy.
- This fundamental difference explains why reconstruction error (Table 1) doesn't guarantee clustering success (Table 2), underscoring the need for imputation objectives to align with downstream analytical goals.

4. **Clustering Impact Complexity:** While imputation accuracy generally correlates with clustering performance, the relationship is not perfectly linear. Statistical analysis indicates that the top-performing methods (Autoencoder, Random Forest, and KNN) achieve comparable clustering effectiveness, as evidenced by overlapping confidence intervals. These findings suggest that preserving the underlying data structure may be as important as minimizing reconstruction error for downstream analytical tasks.
5. **Statistical Reliability:** Bootstrap confidence intervals and non-parametric tests confirm the statistical significance of performance differences, providing robust evidence for method selection recommendations.
6. **Task-Specific Method Selection:** The divergence between imputation accuracy and clustering performance underscores the importance of selecting imputation methods based on specific downstream analytical objectives rather than purely on reconstruction metrics. This finding has important implications for the design of machine learning pipelines in customer analytics applications.
7. **Practical Implications:** Method Selection Framework: The findings establish a clear decision framework for practitioners:
   - **For imputation accuracy,** prioritize Random Forest when downstream applications require precise reconstruction of individual values.
   - **For clustering applications,** consider Autoencoder despite its higher reconstruction error, as it better preserves the feature space structure, which is essential for effective customer segmentation.
   - **For general-purpose applications,** KNN provides balanced performance across both accuracy and clustering metrics.
   - **Method selection should align with downstream analytical objectives** rather than focusing solely on reconstruction metrics.

5. CONCLUSION

5.1 Research Summary

This study aims to compare the performance of multiple missing-data imputation techniques on telecommunications customer data and to evaluate their impact on customer clustering performance. Through systematic comparison of seven imputation methods across three missing mechanisms (MCAR, MAR, MNAR) and four missing rates (5%, 10%, 20%, 30%), several key findings emerged:

5.1.1 Imputation Performance

Random Forest demonstrated superior performance, achieving a Mean Absolute Error (MAE) of 0.1568 and a Root Mean Square Error (RMSE) of 0.2123, outperforming traditional imputation methods (mean/mode) by 53.7

Performance hierarchy from best to worst: Random Forest, Regression/MICE (identical results), KNN, Autoencoder, Mean/Mode, and Forward Fill, respectively.

Robustness across missing mechanisms: Machine learning techniques (Random Forest, MICE, KNN) maintained consistent performance across all missing mechanisms, while traditional methods showed significant degradation under MNAR conditions.

5.1.2 Clustering Impact

Non-linear relationship between imputation quality and clustering performance, with Autoencoder achieving the highest K-Means silhouette score (0.1510) despite moderate imputation accuracy.

K-means proved to be the most suitable clustering method for the telecommunications customer dataset, whereas DBSCAN consistently yielded negative silhouette scores, indicating poor alignment with the data's characteristics.

5.1.3 Statistical Reliability

Statistical significance was confirmed by the Friedman test ($\chi^2 = 55.85$, $p < 0.001$), followed by pairwise Wilcoxon signed-rank tests, which demonstrated the statistically significant superior performance of the Random Forest method.

95% confidence intervals from Bootstrap analysis demonstrated result stability, with Random Forest's MAE confidence interval at [0.1161, 0.1974].

5.2 Discussion

5.2.1 Random Forest Superiority

The superior performance of Random Forest imputation aligns with ensemble learning theory and previous research across multiple dimensions.

From an Ensemble Learning Theory Perspective, Random Forest employs bootstrap aggregating (bagging) principles to reduce variance and prediction error [1]. The bootstrap aggregating mechanism enhances robustness to missing-data uncertainty compared to single-model approaches. Consistent with this observation, prior studies have reported similar effectiveness of Random Forest methods for handling missing values in biological datasets [20].

Consistent with comprehensive surveys of imputation techniques [2], ensemble-based machine learning methods consistently outperform single-model approaches, particularly for complex telecommunications customer data characterized by mixed variable types and non-linear relationships.

5.2.2 MICE and Regression Equivalence

The identical results between MICE and Regression imputation (MAE = 0.2151, RMSE = 0.2643) are an intriguing finding, explicable by MICE's working principles.

MICE Convergence Principles: MICE uses iterative prediction through regression models for each missing variable. When data exhibit stable characteristics and relatively simple missingness patterns, the iterative process converges to solutions comparable to single regression imputation. Consistent with missing-data theory, prior research has shown that under balanced missingness conditions, MICE converges rapidly and yields results similar to those of regression-based methods [14].

Data Quality Validation: This result reflects the characteristics of the Telco Customer Churn dataset, which exhibits relatively straightforward linear relationships among variables, enabling regression-based methods to capture these relationships effectively.

5.2.3 MNAR Challenges

The results indicate that MNAR is the most challenging missing-data mechanism across all evaluated techniques, consistent with theoretical expectations and prior research.

Fundamental MNAR Problems: According to missing-data theory [6], MNAR occurs when missingness depends on the unobserved values themselves, rendering the missing-data mechanism non-identifiable from the observed data. Research on variational autoencoders [17] showed that even advanced techniques like variational autoencoders have limitations in handling MNAR.

Machine Learning Adaptation: Despite the challenges posed by MNAR conditions, machine learning methods such as Random Forest consistently outperformed traditional techniques, owing to their ability to capture complex, non-linear relationships in the data.

5.2.4 Non-linear Relationship Between Imputation Quality and Clustering Performance

The finding that the Autoencoder achieved the highest clustering performance despite moderate imputation accuracy represents a significant discovery consistent with representation learning theory. As demonstrated through confidence interval analysis (Section 4.4.1), while the Autoencoder achieved numerically superior clustering performance (silhouette score: 0.1510), this difference was not statistically significant compared to Random Forest (0.1444), suggesting that multiple imputation approaches may achieve comparable clustering effectiveness through different mechanisms.

The observation that the Autoencoder achieved the highest clustering performance despite moderate imputation accuracy highlights a fundamental distinction between pointwise reconstruction and feature-space preservation.

Unlike Random Forest, which primarily minimizes pointwise prediction errors, autoencoders compress input features into a lower-dimensional latent representation. This bottleneck mechanism compels the model to learn underlying structures and correlations among customer attributes that are critical for clustering algorithms such as K-means when forming well-separated groups. Consequently, autoencoder-based imputation can enhance cluster cohesion even when reconstruction accuracy is not optimal.

The learned latent patterns preserve relationships between service usage, spending behavior, and contract characteristics—patterns that are more important for segmentation than exact value recovery. Conversely, Random Forest excels at predicting each missing point but may not maintain global covariance structure, leading to slightly less coherent clusters despite superior MAE/RMSE performance.

This interpretation is consistent with representation learning theory, which suggests that models trained to encode compressed feature spaces often capture discriminative structure more effectively than models optimized solely for numerical accuracy. The statistical results in Section 4.4.1 further support this observation. Although the Autoencoder achieved the highest silhouette score, its confidence interval substantially overlapped with that of Random Forest, indicating that multiple imputation methods can preserve clustering structure through different underlying mechanisms.

Overall, these findings reinforce that downstream analytical goals—not reconstruction metrics alone—should guide imputation method selection. For clustering-oriented applications, preserving inter-feature relationships may yield greater benefit than optimizing pointwise accuracy.

Although this study provides statistical validation of clustering performance, visual cluster exploration (e.g., PCA or t-SNE projections) was not included due to space and scope limitations. Incorporating these visualization techniques into future work will facilitate a more comprehensive qualitative assessment of the preservation of cluster structure across imputation methods.

5.2.5 Research Objectives Support

These findings comprehensively support all four research objectives:

Objective 1: Performance comparisons are supported by a comprehensive evaluation across 84 experimental conditions, with statistical tests confirming the results.
Objective 2: Customer clustering models demonstrate high accuracy, as measured by silhouette scores and complementary clustering metrics.
Objective 3: The impact analysis of missing-data types and missingness rates revealed apparent differences among the MCAR, MAR, and MNAR mechanisms.
Objective 4: This study derives guidelines for technique selection from empirical evidence and rigorous statistical analysis.

5.3 Research Limitations

This research has several limitations requiring consideration in result interpretation:

1. **Single-Dataset Limitation:** This study relies on a single dataset, namely the IBM Telco Customer Churn dataset, a widely recognized benchmark in telecommunications churn prediction research. While its extensive use in prior studies supports its validity, reliance on a single dataset may limit the generalizability of the findings to other telecommunications customer datasets or different industry contexts. The dataset reflects customer demographics, service usage patterns, and churn behaviors specific to a particular telecommunications provider, which may not capture the diversity of customer bases across markets, regions, or service types. Moreover, customer churn dynamics can vary substantially across telecommunications companies due to differences in market maturity, competitive conditions, regulatory environments, and cultural factors. Future research should therefore validate the proposed findings using multiple datasets from diverse telecommunications providers and geographical regions, and potentially from other subscription-based industries, to assess the robustness and broader applicability of the observed missing-data handling patterns.
2. **Simulated Missing Data:** This study relies on simulated missingness under three mechanisms (MCAR, MAR, and MNAR) rather than real-world missing data. Although simulation enables controlled experimentation and reproducibility, it may not fully capture the complexity of missing data patterns observed in practical business settings, where observed or unobserved factors can systematically influence missingness.
3. **Limited Clustering Evaluation Methodology:** The clustering performance assessment relied primarily on silhouette scores as the evaluation metric, which may not provide comprehensive insight into clustering quality. This limitation manifests in several ways:
   - **Single Metric Dependency:** Using only silhouette scores may miss other essential clustering characteristics that alternative metrics could capture.
   - **Missing Complementary Metrics:** Future work should include additional clustering validation indices, such as the Calinski-Harabasz Index (which measures the between-cluster-to-within-cluster variance ratio) and the Davies-Bouldin Index (which evaluates cluster separation and compactness), to provide a more robust assessment of clustering quality.
   - **Lack of Visual Validation:** Visual cluster validation using dimensionality reduction techniques, such as t-SNE or PCA, provides valuable qualitative insight into the preservation of clustering structure across different imputation methods.
   - **Cluster Stability Analysis:** The study did not assess cluster stability across multiple runs with different random initializations, which is essential for robustness validation and for ensuring that observed performance differences are not artifacts of specific algorithmic initializations.

   Silhouette score was selected as the primary clustering evaluation metric in this study due to its robustness and independence from ground-truth cluster labels, making it well-suited for unsupervised customer segmentation where no predefined classes exist. The metric simultaneously evaluates intra-cluster cohesion and inter-cluster separation using pairwise distances, making it suitable for examining how different imputation methods preserve the underlying feature-space structure. Moreover, the Silhouette score remains stable across a wide range of cluster shapes and sample sizes, making it a reliable standalone indicator of clustering quality in customer analytics applications.

   Nevertheless, we acknowledge that relying solely on the Silhouette score may not capture all nuances of clustering behavior. Future studies should incorporate complementary validity indices, such as the Calinski–Harabasz and Davies–Bouldin scores, to enable a multi-perspective evaluation of clustering compactness and separability.
4. **Technique Scope and Method Selection Limitations:** This study was limited to seven established missing-data handling techniques, representing commonly used approaches in the field. However, more advanced methods—such as deep learning–based imputation, matrix factorization techniques, and ensemble imputation frameworks—may offer superior performance

and warrant further investigation in future studies [2].

Additional Method Selection Scope Limitation: While our seven-method selection provides comprehensive coverage across traditional, machine learning, and deep learning approaches, the exclusion of advanced techniques like GAIN [45], VAE [46], and Transformer-based methods [39] represents a significant limitation. These exclusions reflect computational feasibility and practical deployment constraints in telecommunications environments. Future research that evaluates these advanced methods in telecommunications settings could yield valuable methodological insights, particularly for organizations with sufficient computational resources and technical expertise to implement complex imputation frameworks. As generative AI and attention-based models continue to advance, future research should validate the present findings against these emerging approaches as computational barriers decrease.

5. **Short-term Evaluation:** This study focused primarily on model-level performance metrics and did not assess the long-term business impacts of deploying different missing-data handling techniques, including effects on customer retention, revenue outcomes, or operational feasibility in production environments.

Despite these limitations, the qualitative and statistical evidence consistently indicate that representation-preserving imputers can outperform accuracy-focused methods in clustering tasks.

5.4 Recommendations

5.4.1 Practical Recommendations

Technique Selection by Scenario:

1. **Low Missing Data (< 10%):** Apply Random Forest imputation to achieve robust, stable performance across different missing-data mechanisms, as it effectively preserves the data structure under low missingness.
2. **High Missing Data (> 20%):** Employ advanced imputation techniques such as Random Forest or deep learning–based approaches (e.g., Autoencoders) to preserve underlying data structures when substantial proportions of values are missing.
3. **Mixed-type Data:** Recommend MICE or Random Forest for handling both continuous and categorical variables.
4. **Outlier Detection and Management:** Implement before normalizing error metrics to prevent erroneous evaluation.

Organizational Implementation:

1. **Automated Pipeline Creation:** For missing pattern detection and appropriate technique selection.
2. **Data Scientist Training:** Ensure understanding of technique principles and limitations.
3. **Continuous Model Performance Monitoring:** Establish continuous monitoring mechanisms to assess the effectiveness of imputation techniques using reconstruction error metrics and downstream model performance, ensuring long-term reliability under evolving data conditions.

5.4.2 Future Research Recommendations

New Technique Development:

1. **Hybrid Imputation Methods:** Combining multiple technique strengths, such as Random Forest for initial imputation with deep learning fine-tuning.
2. **Domain-specific Imputation:** Considering telecommunications data characteristics like temporal patterns and customer lifecycle.
3. **Federated Imputation:** Managing multi-source customer data without privacy violations.

Expanded Study Scope:

1. **Time Series Data:** Studying performance in temporal customer data, like usage patterns over time.
2. **Multi-domain Validation:** Testing with data from other industries like banking, e-commerce, and healthcare.
3. **Real-world Implementation Studies:** Long-term impact studies in actual business environments.
4. **Explainable AI Integration:** Developing interpretable tools for imputation technique selection decisions.

Business Impact Assessment:

1. **Customer Lifetime Value (CLV) Impact:** Studying imputation quality effects on CLV prediction.
2. **Marketing Campaign Effectiveness:** Evaluating how improved customer clustering affects marketing campaign effectiveness.
3. **Cost-Benefit Analysis:** Analyzing investment costs and returns for complex imputation systems.

5.4.3 Expected Business Impact

Improved Customer Clustering Accuracy through Random Forest imputation with 53.7% lower MAE than traditional methods will enhance clustering accuracy, resulting in:

1. **More Effective Marketing Strategies** through more profound, more accurate customer group understanding.
2. **Reduced Customer Retention Costs** via risk group identification and appropriate strategy adjustment.

3. **Increased Customer Lifetime Value** through targeted product and service offerings matching customer needs.

5.5 Discussion in Relation to Prior Telco Churn Studies

Several prior studies have employed the Telco Customer Churn Dataset primarily to evaluate predictive performance using various classification models [64]. These works typically focus on optimizing churn prediction accuracy under fixed preprocessing and complete-data assumptions.

In contrast, the objective of this study is not to compete with existing prediction-oriented approaches in terms of classification performance. Instead, we investigate the impact of different missing-data imputation techniques on the preservation of data structure and clustering behavior. Our findings show that imputation methods achieving strong reconstruction or prediction performance do not necessarily preserve meaningful clustering structures, revealing an imputation–clustering dissociation that prior churn prediction studies have overlooked.

Therefore, this study contributes to the existing literature by offering a data-centric perspective on preprocessing choices, emphasizing their impact on data integrity and downstream analytical outcomes, rather than positioning itself as a replacement for established predictive models. This perspective is particularly valuable for exploratory analysis and clustering-based applications, where preserving underlying data structures is essential for deriving meaningful insights.

5.6 Conclusion

This research demonstrates that the choice of missing-data imputation techniques has a substantial impact on customer clustering performance in telecommunications analytics. Random Forest imputation consistently outperforms other methods under low-to-moderate levels of missing data.

These findings not only reinforce ensemble learning theory and prior evidence on the effectiveness of machine learning methods for handling missing data but also offer new insights into the complex relationship between imputation quality and downstream analytical performance.

The practical recommendations derived from this research can guide the development of effective customer analytics systems, enabling organizations to utilize customer data more effectively despite incomplete information. As a result, organizations can make more accurate and timely business decisions, strengthening their competitive position in the telecommunications market.

AUTHOR CONTRIBUTIONS

Conceptualization, P.S.; methodology, P.S.; software, P.S.; validation, P.C.; formal analysis, P.C.; investigation, P.C.; data curation, P.S.; writing—original draft preparation, P.S.; writing—review and editing, P.C.; visualization, P.S.; supervision, P.C.; All authors have read and agreed to the published version of the manuscript.

References

[1] Y. Chen, Y. Lv and F. -Y. Wang, "Traffic Flow Imputation Using Parallel Data and Generative Adversarial Networks," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1624-1630, April 2020.

[2] X. Miao, Y. Wu, L. Chen, Y. Gao and J. Yin, "An Experimental Survey of Missing Data Imputation Algorithms," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6630-6650, 1 July 2023.

[3] X. Xu, W. Chong, S. Li, A. Arabo and J. Xiao, "MIAEC: Missing Data Imputation Based on the Evidence Chain," in *IEEE Access*, vol. 6, pp. 12983-12992, 2018.

[4] R. Wu, S. D. Hamshaw, L. Yang, D. W. Kincaid, R. Etheridge and A. Ghasemkhani, "Data Imputation for Multivariate Time Series Sensor Data With Large Gaps of Missing Data," in *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10671-10683, 1 June1, 2022.

[5] Y. Liu, T. Dillon, W. Yu, W. Rahayu and F. Mostafa, "Missing Value Imputation for Industrial IoT Sensor Data With Large Gaps," in *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6855-6867, Aug. 2020.

[6] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] R. Polikar, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, Third Quarter 2006.

[9] Y. Wu, J. Wang, X. Miao, W. Wang and J. Yin, "Differentiable and Scalable Generative Adversarial Models for Data Imputation," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 2, pp. 490-503, Feb. 2024.

[10] J. Venugopalan, N. Chanani, K. Maher and M. D. Wang, "Novel Data Imputation for Multiple Types of Missing Data in Intensive Care Units," in *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1243-1250, May 2019.

[11] W. Khan *et al.*, "Mixed Data Imputation Using Generative Adversarial Networks," in *IEEE Access*, vol. 10, pp. 124475-124490, 2022.

[12] S. E. Awan *et al.*, "Imputation of missing data with class imbalance using conditional genera-

tive adversarial networks," *Neurocomputing*, vol. 453, pp. 164–171, 2021.

[13] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[14] S. Van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2018.

[15] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. London, U.K.: Chapman & Hall/CRC, 1997.

[16] V. Kumar and W. Reinartz, "Creating enduring customer value," *Journal of Marketing*, vol. 80, no. 6, pp. 36–68, 2016.

[17] S. A. Neslin *et al.*, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research*, vol. 43, no. 2, pp. 204–211, 2006.

[18] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, 2nd ed. Boston, MA, USA: Kluwer, 2000.

[19] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[21] Y. Gong, Z. Li, J. Zhang, W. Liu, Y. Yin and Y. Zheng, "Missing Value Imputation for Multi-View Urban Statistical Data via Spatial Correlation Learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 686-698, 1 Jan. 2023.

[22] P. Wang, T. Hu, F. Gao, R. Wu, W. Guo and X. Zhu, "A Hybrid Data-Driven Framework for Spatiotemporal Traffic Flow Data Imputation," in *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 16343-16352, 1 Sept.1, 2022.

[23] Y. Zelenkov and A. Suchkova, "Predicting customer churn based on changes in their behavior patterns," *Business Informatics*, vol. 17, pp. 7–17, 2023.

[24] A. Chadaga, M. Legg and C. H. B. Liu, "Enhancing customer lifetime value using data science and predictive modeling," *Technium Business and Management*, vol. 12, pp. 112-125, 2025.

[25] H. Li, Y. Liao, Z. Tian, Z. Liu, J. Liu and X. Liu, "Bidirectional Stackable Recurrent Generative Adversarial Imputation Network for Specific Emitter Missing Data Imputation," in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2967-2980, 2024.

[26] X. Chen, M. Lei, N. Saunier and L. Sun, "Low-Rank Autoregressive Tensor Completion for Spatiotemporal Traffic Data Imputation," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12301-12310, Aug. 2022.

[27] M. S. Osman, A. M. Abu-Mahfouz and P. R. Page, "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case," in *IEEE Access*, vol. 6, pp. 63279-63291, 2018.

[28] R. C. Pereira, P. H. Abreu and P. P. Rodrigues, "Partial Multiple Imputation With Variational Autoencoders: Tackling Not at Randomness in Healthcare Data," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4218-4227, Aug. 2022.

[29] X. Zhu, J. Yang, C. Zhang and S. Zhang, "Efficient Utilization of Missing Data in Cost-Sensitive Learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2425-2436, 1 June 2021.

[30] M. J. Kim and Y. Cho, "Imputation of missing values in well log data using k-nearest neighbor collaborative filtering," *Computers & Geosciences*, vol. 193, p. 105712, 2024.

[31] X. Wei, Y. Zhang, S. Wang, X. Zhao, Y. Hu and B. Yin, "Self-Attention Graph Convolution Imputation Network for Spatio-Temporal Traffic Data," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19549-19562, Dec. 2024.

[32] N. Abiri, B. Linse, P. Edén and M. B.O. Ohlsson, "Establishing strong imputation performance of a denoising autoencoder," *Neurocomputing*, vol. 365, pp. 137–146, 2019.

[33] R. C. Pereira, P. H. Abreu and P. P. Rodrigues, "Siamese autoencoder architecture for the imputation of data missing not at random," *Journal of Computational Science*, vol. 78, p. 102269, 2024.

[34] X. Lai, X. Wu, and L. Zhang, "Autoencoder-based multi-task learning for imputation and classification," *Applied Soft Computing*, vol. 98, p. 106838, 2021.

[35] R. Shahbazian and S. Greco, "Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey and Evaluation," in *IEEE Access*, vol. 11, pp. 88908-88928, 2023.

[36] S. E. Awan *et al.*, "Imputation of missing data with class imbalance using conditional generative adversarial networks," *Neurocomputing*, vol. 453, pp. 164–171, 2021.

[37] Z. Guo, Y. Wan and H. Ye, "A data imputation method for multivariate time series based on GAN," *Neurocomputing*, vol. 360, pp. 185–197, 2019.

[38] J. Zhao, C. Rong, C. Lin and X. Dang, "Multivariate time series data imputation using attention-based mechanism," *Neurocomputing*, vol. 542, p. 126238, 2023.

[39] D. Liu, Y. Wang, C. Liu, K. Wang, X. Yuan and

C. Yang, "Blackout Missing Data Recovery in Industrial Time Series Based on Masked-Former Hierarchical Imputation Framework," in *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 1138-1150, April 2024.

[40] N. Karmitsa, S. Taheri, A. Bagirov and P. Mäkinen, "Missing Value Imputation via Clusterwise Linear Regression," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1889-1901, 1 April 2022.

[41] A. A. Harder, G. R. Olbricht, G. Ekuma, D. B. Hier and T. Obafemi-Ajayi, "Multiple Imputation for Robust Cluster Analysis to Address Missingness in Medical Data," in *IEEE Access*, vol. 12, pp. 42974-42991, 2024.

[42] L. Zhao, Z. Chen, Z. Yang, Y. Hu and M. S. Obaidat, "Local Similarity Imputation Based on Fast Clustering for Incomplete Data in Cyber-Physical Systems," in *IEEE Systems Journal*, vol. 12, no. 2, pp. 1610-1620, June 2018.

[43] A. Tharwat and W. Schenck, "Active Learning for Handling Missing Data," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3273-3287, Feb. 2025.

[44] "Telco Customer Churn Dataset," *Kaggle*, 2023. [Online]. Available: `https://www.kaggle.com/datasets/jethwaaatmik/telco-customer-churn-dataset`

[45] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 5689–5698, 2018.

[46] A. Nazábal, P. M. Olmos, Z. Ghahramani and I. Valera, "Handling incomplete heterogeneous data using VAEs," *Pattern Recognition*, vol. 107, p. 107501, 2020.

[47] D. F. Heitjan and S. Basu, "Distinguishing "missing at random" and "missing completely at random"," *The American Statistician*, vol. 50, no. 3, pp. 207–213, 1996.

[48] Y. He , A. M. Zaslavsky, D. P. Harrington, P. Catalano and M. B. Landrum , "Multiple imputation in a large-scale complex survey," *Stat. Methods Med. Res.*, vol. 19, no. 6, pp. 653–670, 2010.

[49] K. Potdar, T. S. Pardawala and C. D. Pai, "A comparative study of categorical variable encoding techniques," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, 2017.

[50] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data," *Journal of Big Data*, vol. 7, no. 1, pp. 1–45, 2020.

[51] P. Cerda, G. Varoquaux and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Machine Learning*, vol. 107, no. 8, pp. 1477–1494, 2018.

[52] F. Pargent, F. Pfisterer, J. Thomas and B. Bischl , "Regularized target encoding outperforms traditional methods," *Computational Statistics* , vol. 37, no. 5, pp. 2671–2692, 2022.

[53] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[54] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Arti cial Intelligence and Statistics (AISTATS)*, vol. 15, pp. 315–323, 2011.

[55] N. Srivastava, G. Hintonm, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint* arXiv:1412.6980, 2014.

[57] P. Vincent *et al.*, "Stacked denoising autoencoders," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[58] P. Baldi and K. Hornik, "Neural networks and principal component analysis," *Neural Networks*, vol. 2, no. 1, pp. 53–58, 1989.

[59] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Stat. Anal. Data Min.*, vol. 10, no. 6, pp. 363–377, 2017.

[60] J. Josse and F. Husson, "missMDA: A package for handling missing values," *Journal of Statistical Software*, vol. 70, no. 1, pp. 1–31, 2016.

[61] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[62] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[63] U. von Luxburg, "Clustering stability: An overview," *Foundations and Trends® in Machine Learning*, vol. 2, no. 3, pp. 235–274, 2010.

[64] Y. Zhang, "Machine learning-based prediction of telecom customer churn," *Journal of Science and Technology*, vol. 18, no. 2, pp. 116–123, 2025.

Patthama Sukthong received her B.S. and M.S. degrees in Computer Science from Thammasat University, Thailand. She is currently pursuing a doctoral degree in Computer Science at King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand. She is also working in the Business Process Management Division at National Telecom Public Company Limited. Her research interests include data mining and machine learning.

Pattama Charoenporn received her B.S. degree in Computer Science from Thammasat University, Thailand, and her M.S. degree in Business Software Development from Chulalongkorn University, Thailand. She also received a postgraduate degree in Computer Science from King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand. She is currently an Associate Professor in Computer Science at KMITL, Thailand. She can be contacted at pattama.ch@kmitl.ac.th.